Data preprocessing and integration for reproducible multiomic biomarker discovery and validation

W. Evan Johnson

Division of Computational Biomedicine
Boston University School of Medicine
wej@bu.edu

http://jlab.bu.edu/
@wejlab

April 29, 2013

Four Steps to Biomarker Validation

Optimal Biomarker Generation and Validation:

1. Carefully designed study (begin with end in mind):
   - Targeted data collection
   - Include 'orthogonal' data types

Four Steps to Biomarker Validation

Optimal Biomarker Generation and Validation:

1. Carefully designed study (begin with end in mind):
   - Targeted data collection
   - Include 'orthogonal' data types
2. Robust development of reproducible biomarker:
   - Data preprocessing and integration
   - Discovery, development, optimization (and adaptation!)

Four Steps to Biomarker Validation

Optimal Biomarker Generation and Validation:

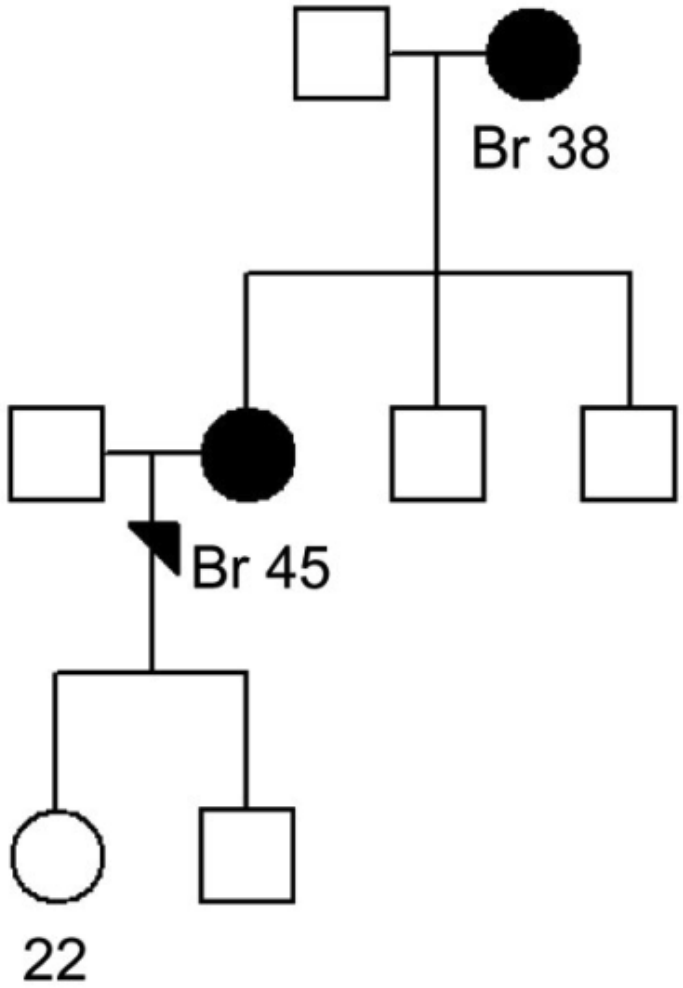
1. Carefully designed study (begin with end in mind):
   - Targeted data collection
   - Include 'orthogonal' data types
2. Robust development of reproducible biomarker:
   - Data preprocessing and integration
   - Discovery, development, optimization (and adaptation!)
3. Validation types (predefined if possible):
   - Internal and external validation
   - Mechanistic or functional validation
   - Additional genomic profiling?

Four Steps to Biomarker Validation

Optimal Biomarker Generation and Validation:

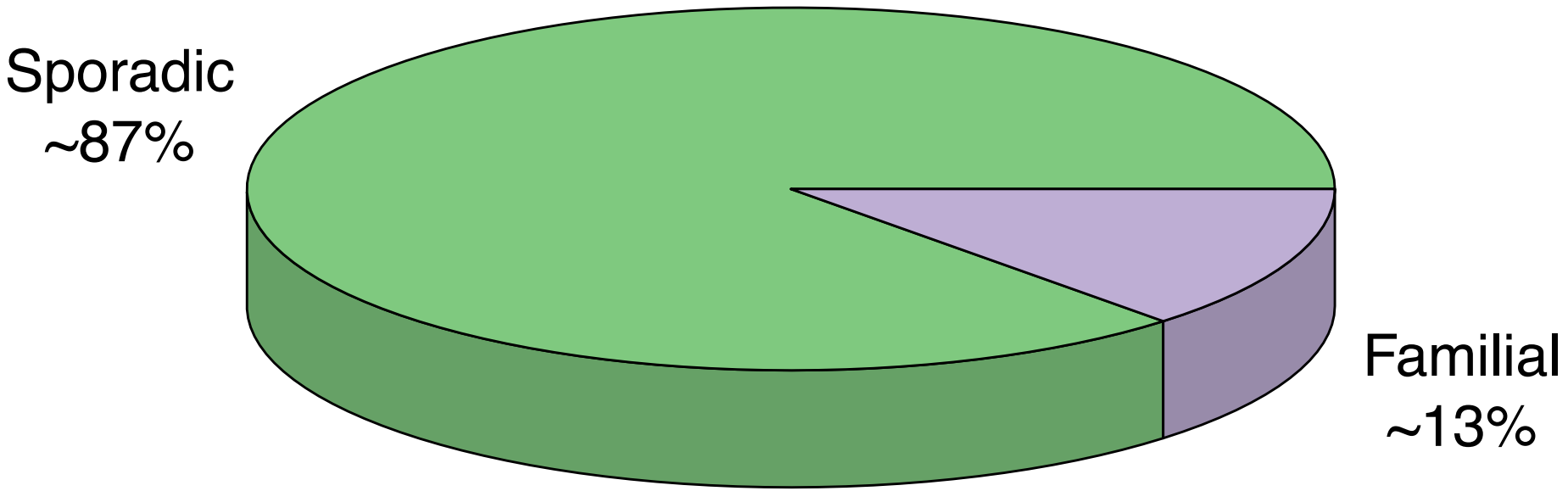
1. Carefully designed study (begin with end in mind):
   - Targeted data collection
   - Include 'orthogonal' data types
2. Robust development of reproducible biomarker:
   - Data preprocessing and integration
   - Discovery, development, optimization (and adaptation!)
3. Validation types (predefined if possible):
   - Internal and external validation
   - Mechanistic or functional validation
   - Additional genomic profiling?
4. Last but not least: Transparency!

Breast Cancer Often Runs in Families



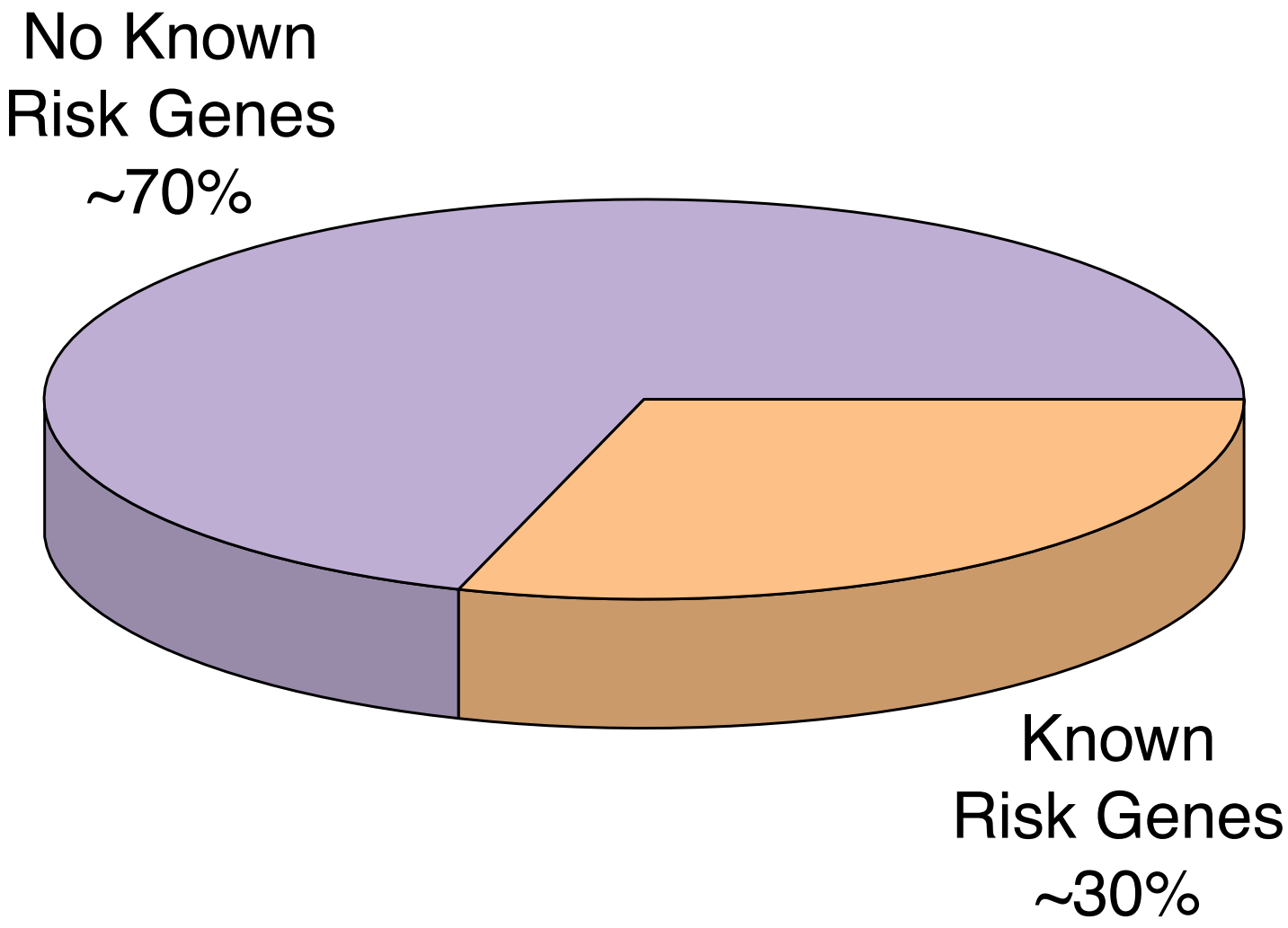
http://openi.nlm.nih.gov

Breast Cancer



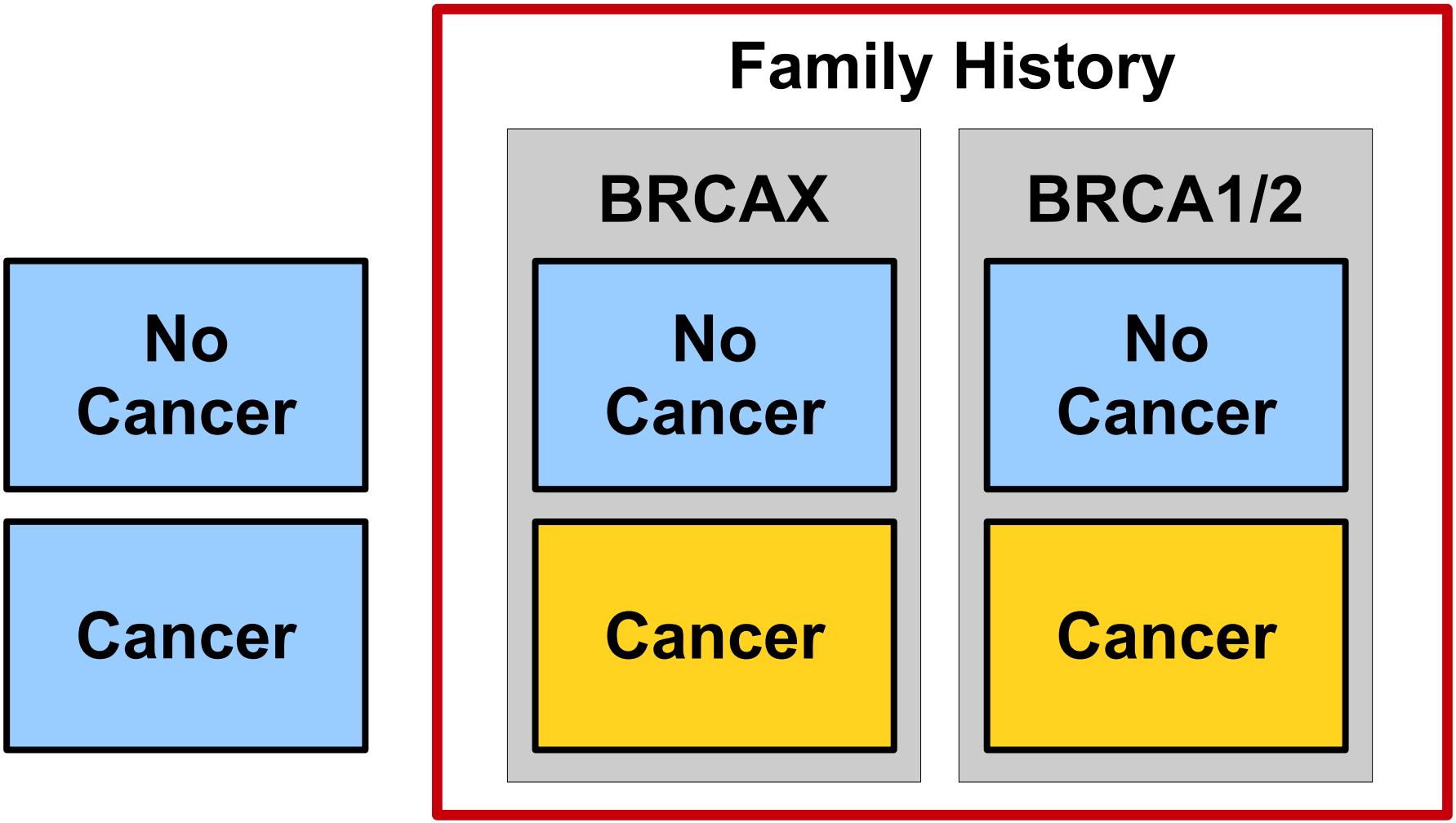
Collaborative Group on Hormonal Factors in Breast Cancer

Familial Breast Cancer



Stratton and Rahman, *Nature Genetics*, 2008

Risk Subpopulations



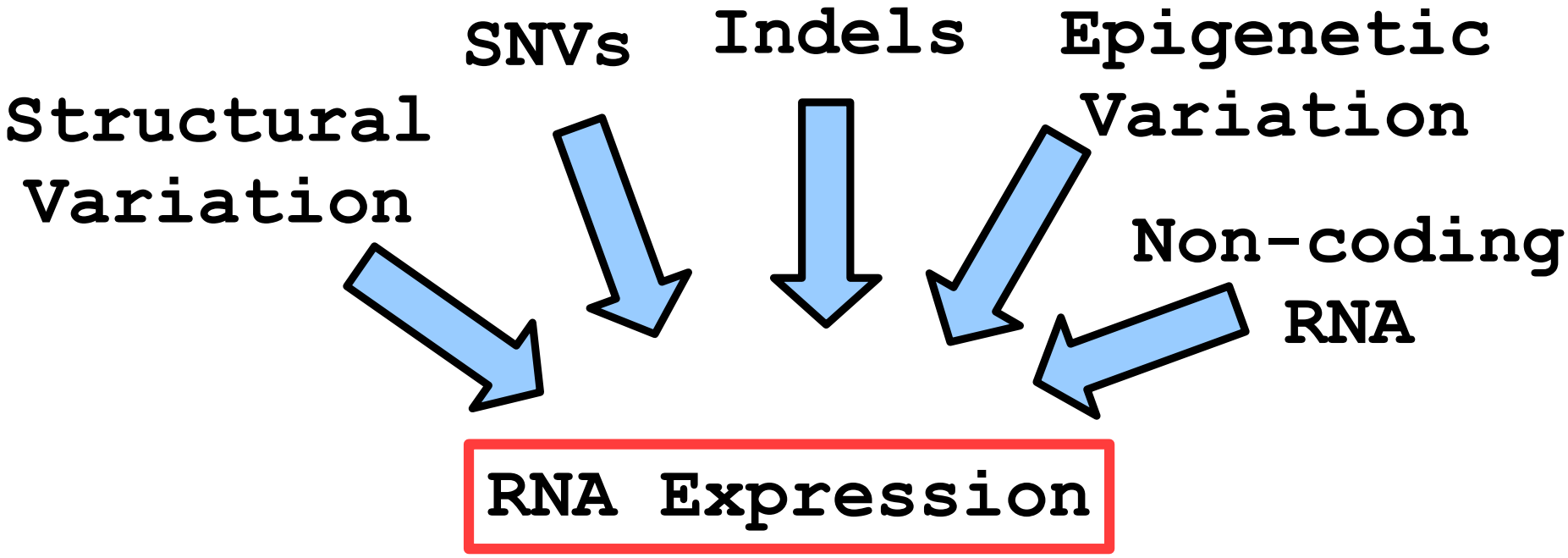
Predicting Hereditary Breast Cancer

Goal: Develop a non-invasive biomarker for hereditary breast cancer risk

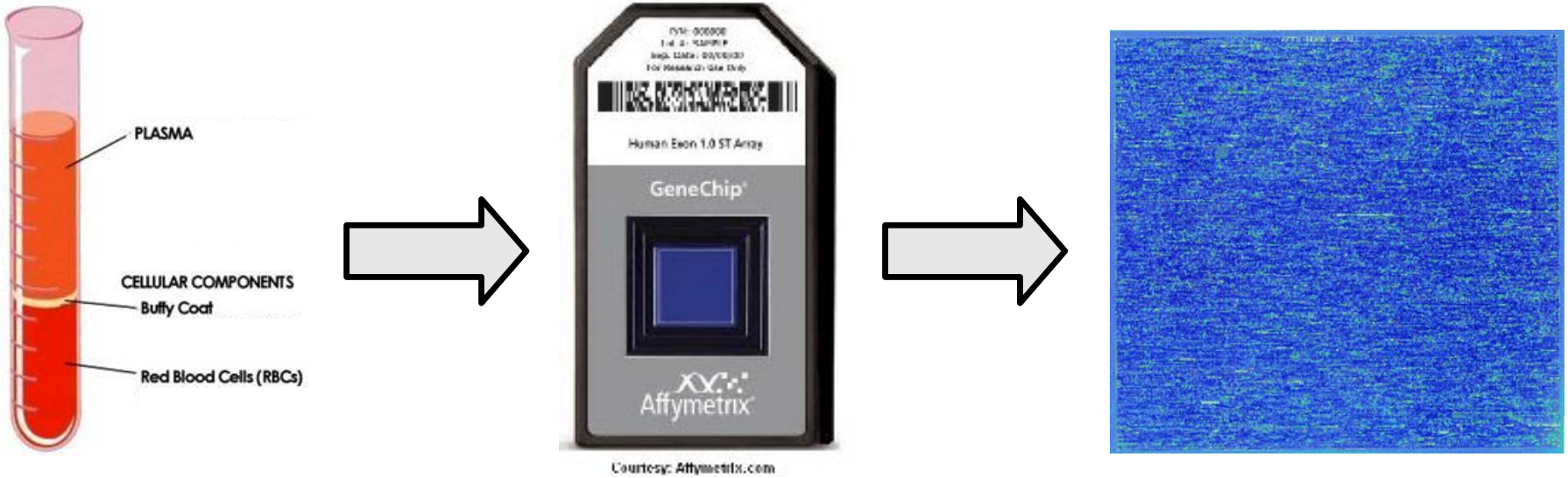
Key hypotheses or information to build our biomarker:

- Rare or highly penetrant genetic variation impacting key pathways (e.g. DNA repair)
- Germline-driven mRNA degregulation as an intermediate risk phenotype for familial breast-cancer susceptibility
- Expression patterns in peripheral blood will enable prospective identification of high-risk women who will develop breast cancer

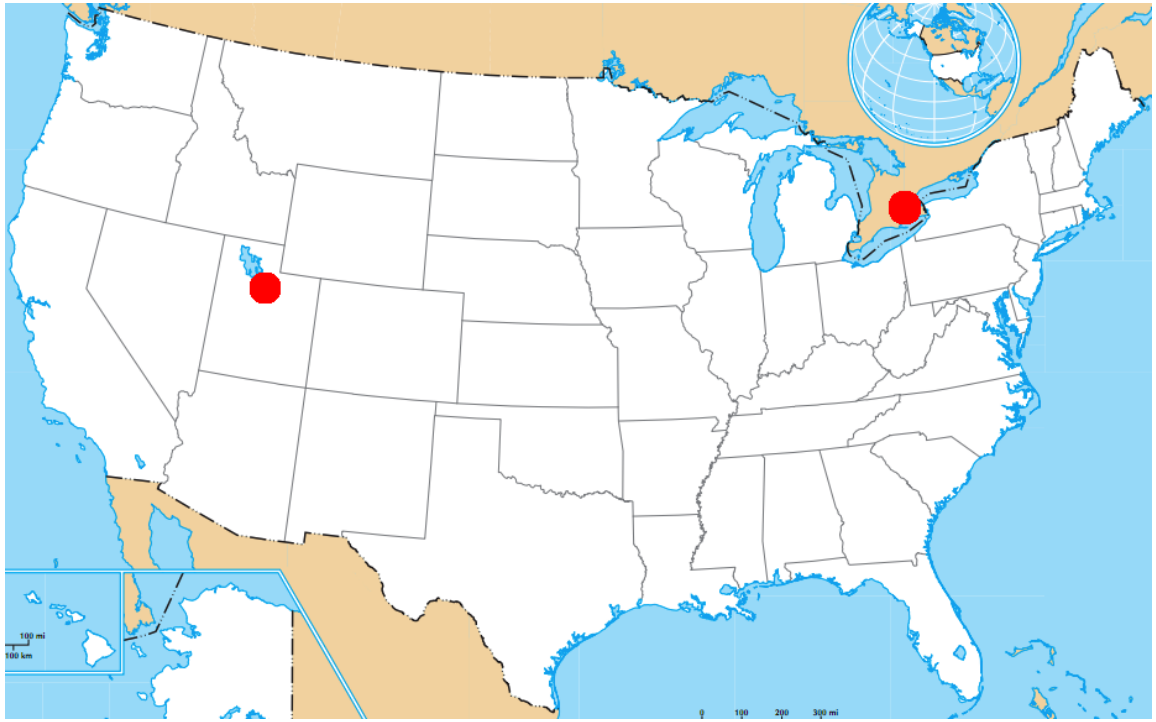
RNA Expression as a Variation Surrogate



Microarray Profiles of Peripheral Blood



Patient Cohorts



Location	Type	Patients
Utah	Retrospective	124
Ontario	Retrospective	36
Ontario	Prospective	37

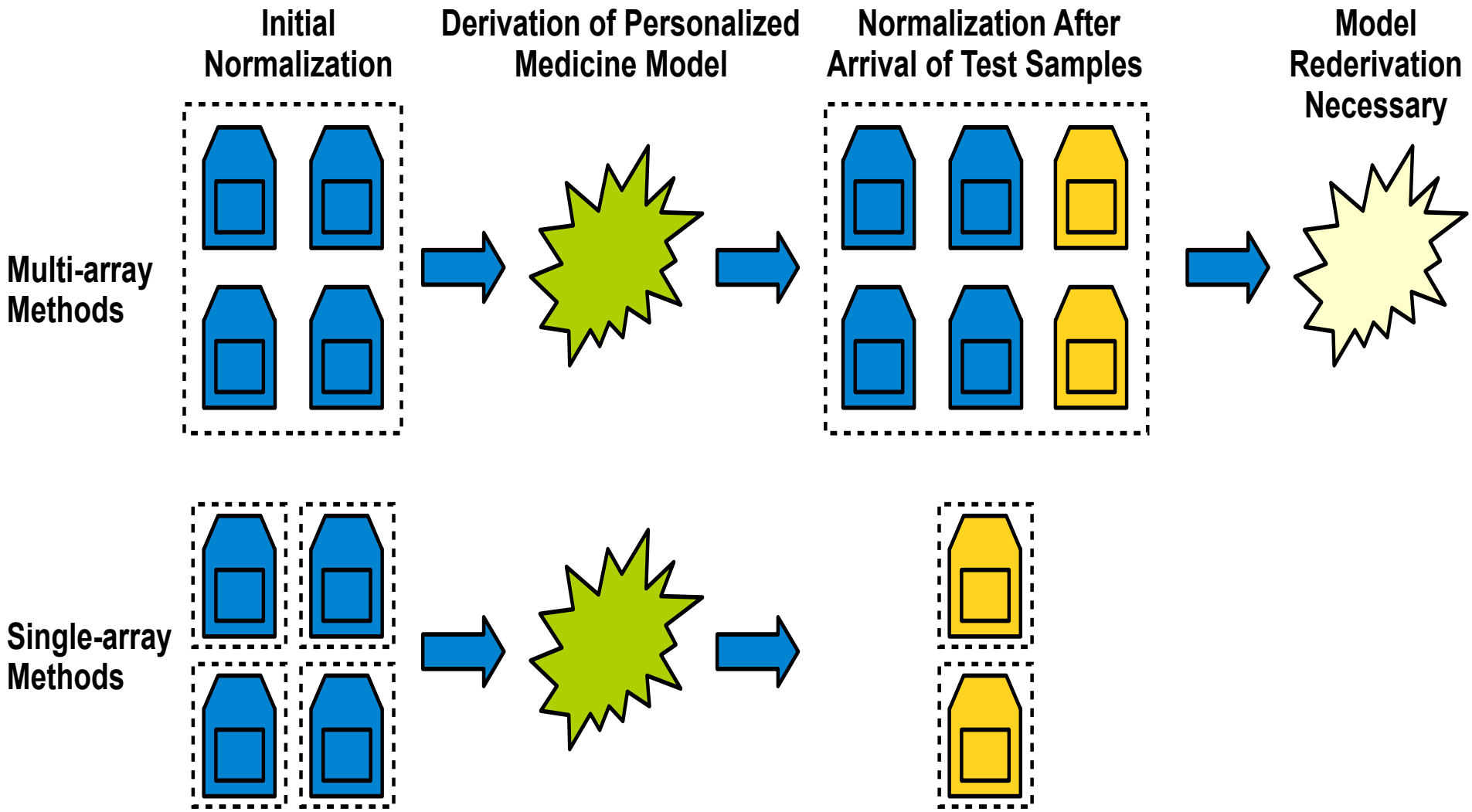
Study Design Summary

Important elements of the study design:

- Study patients recruited and data collected specifically for this study
- Validation on a heterogenous population/dataset
- Clear purpose for the study and 'success' is predefined
- Collected expression array and DNA sequencing data (validation? multi-omic biomarker?)

Goals of Personalized Genomic Medicine

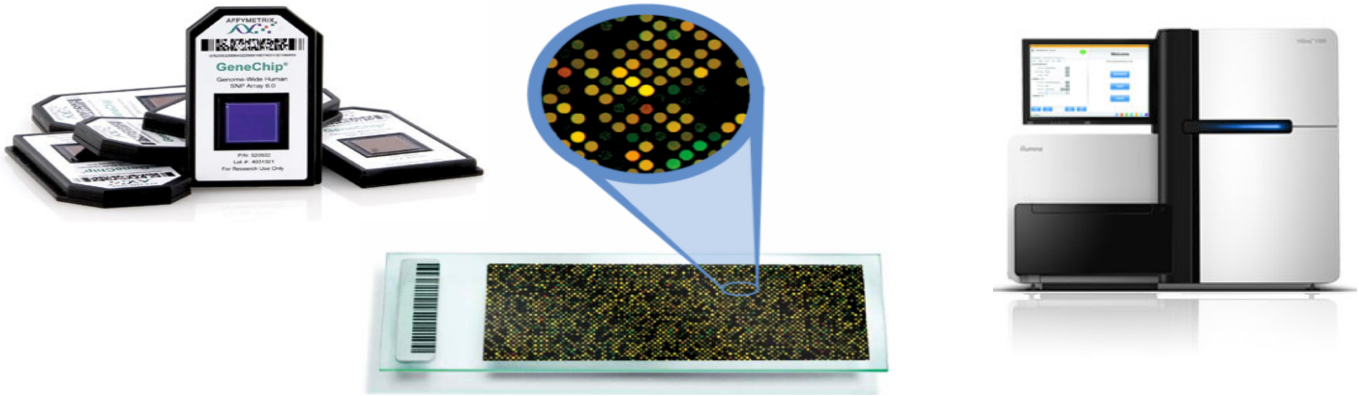
Data Normalization Methods



Microarray/Sequencing Normalization Methods

- Multisample Methods:
  - dChip *(Li, PNAS, 2001)*
  - QQ/RMA *(Bolstad, Bioinformatics, 2003; Irizarry, Biostatistics, 2003)*
  - RNA-seq: Conditional-QQ (Hansen, Biostatistics, 2012)
- Single Sample Methods:
  - MAS5 *(Hubbell, Bioinformatics, 2002)*
  - fRMA *(McCall, Biostatistics, 2010)*
  - Barcoding *(Irizarry Nat Meth 2009; McCall, NAR, 2012)*
  - RPKM/FPKM *(Mortazavi, Nat Meth, 2008; Trapnell, Nat Biotech 2010)*
- **Single Array and Sequencing Integrative Methods:**
  - **SCAN-UPC** *(Piccolo, Genomics, 2012)*

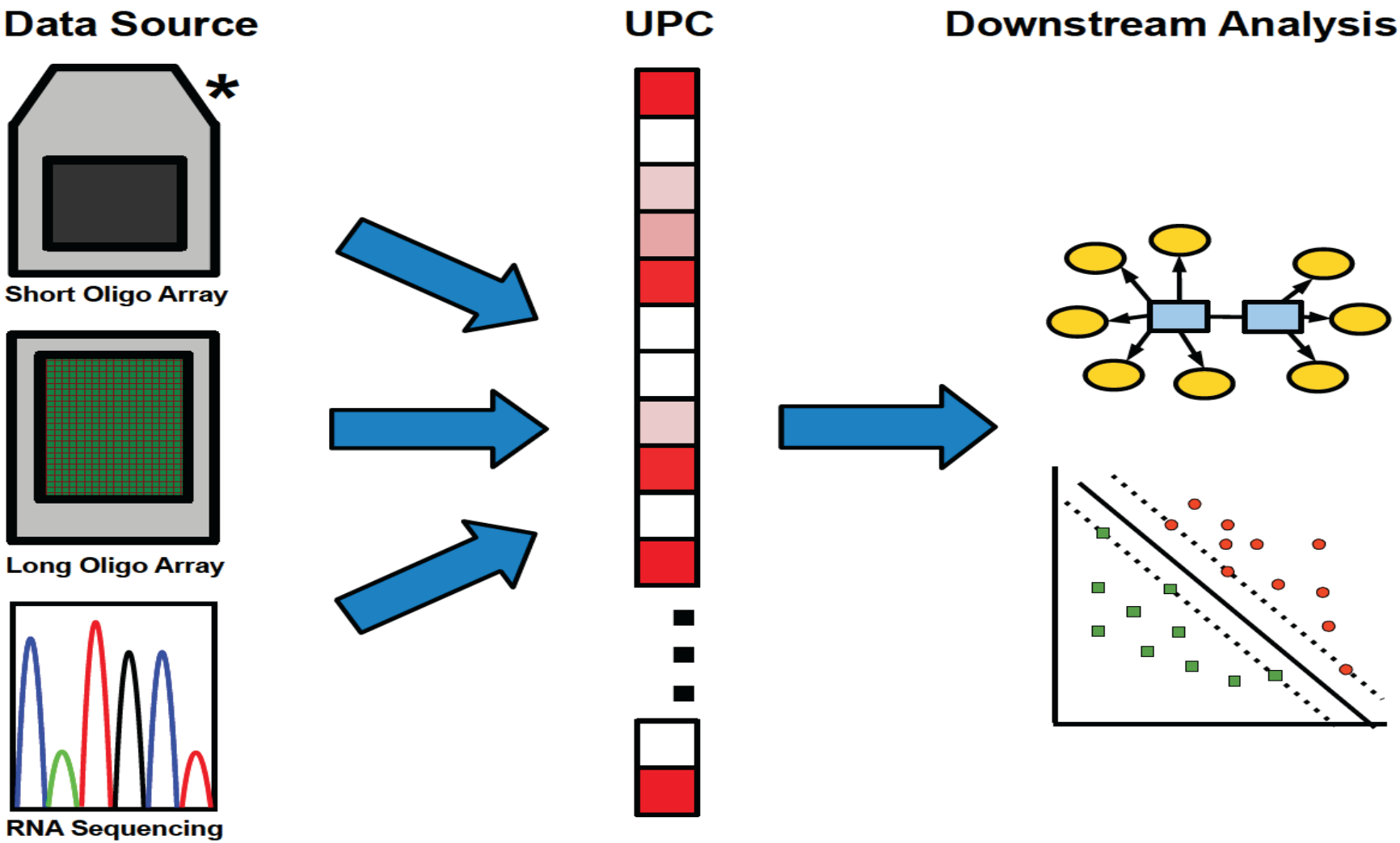
SCAN-UPC: Single Sample of Expression Estimates



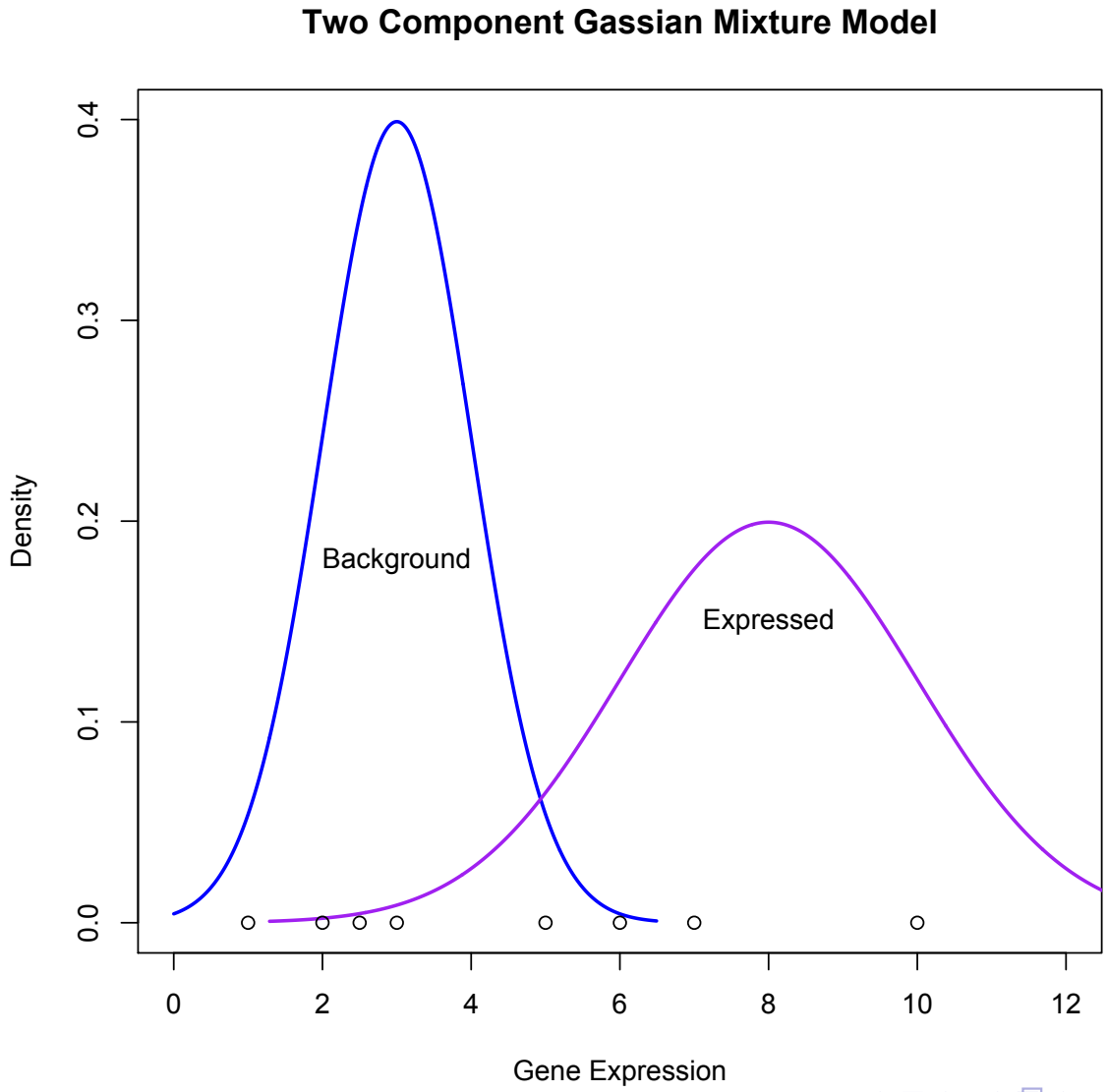
Our approaches: Single Channel Array Normalization (SCAN), and Universal Probability of Expression Codes (UPC)

- Uses background *within* a single sample
- Individual patient samples without any extraneous data
- Can be applied to all platforms: one and two color arrays, RNA-seq (UPC)
- Naturally combines data across platforms (UPC)

Biomarker and Personalized Medicine Workflows

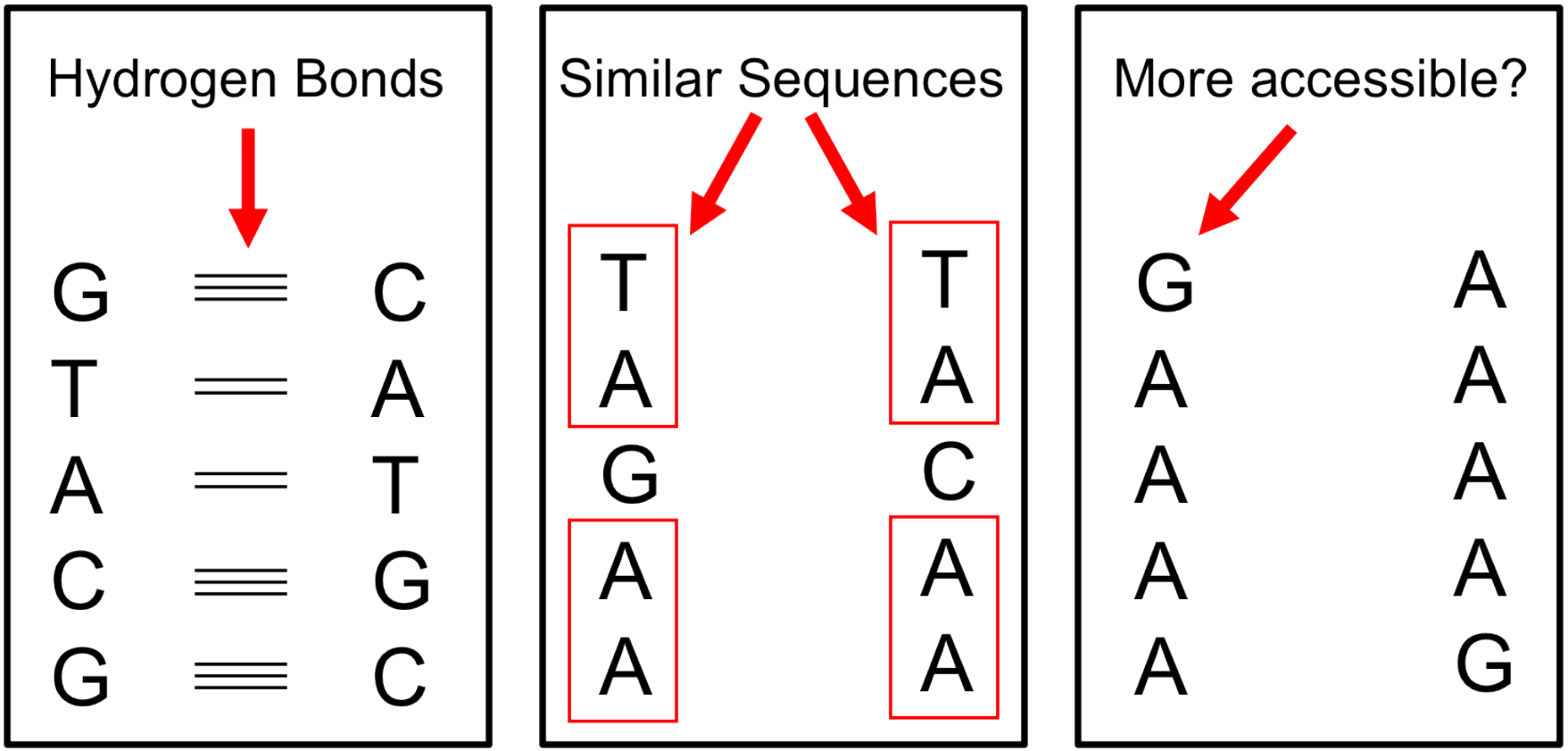


SCAN-UPC: Single Sample of Expression Estimates



Universal Probability of Expression Code (UPC)

SCAN-UPC Model Justification:



Universal Probability of Expression Code (UPC)

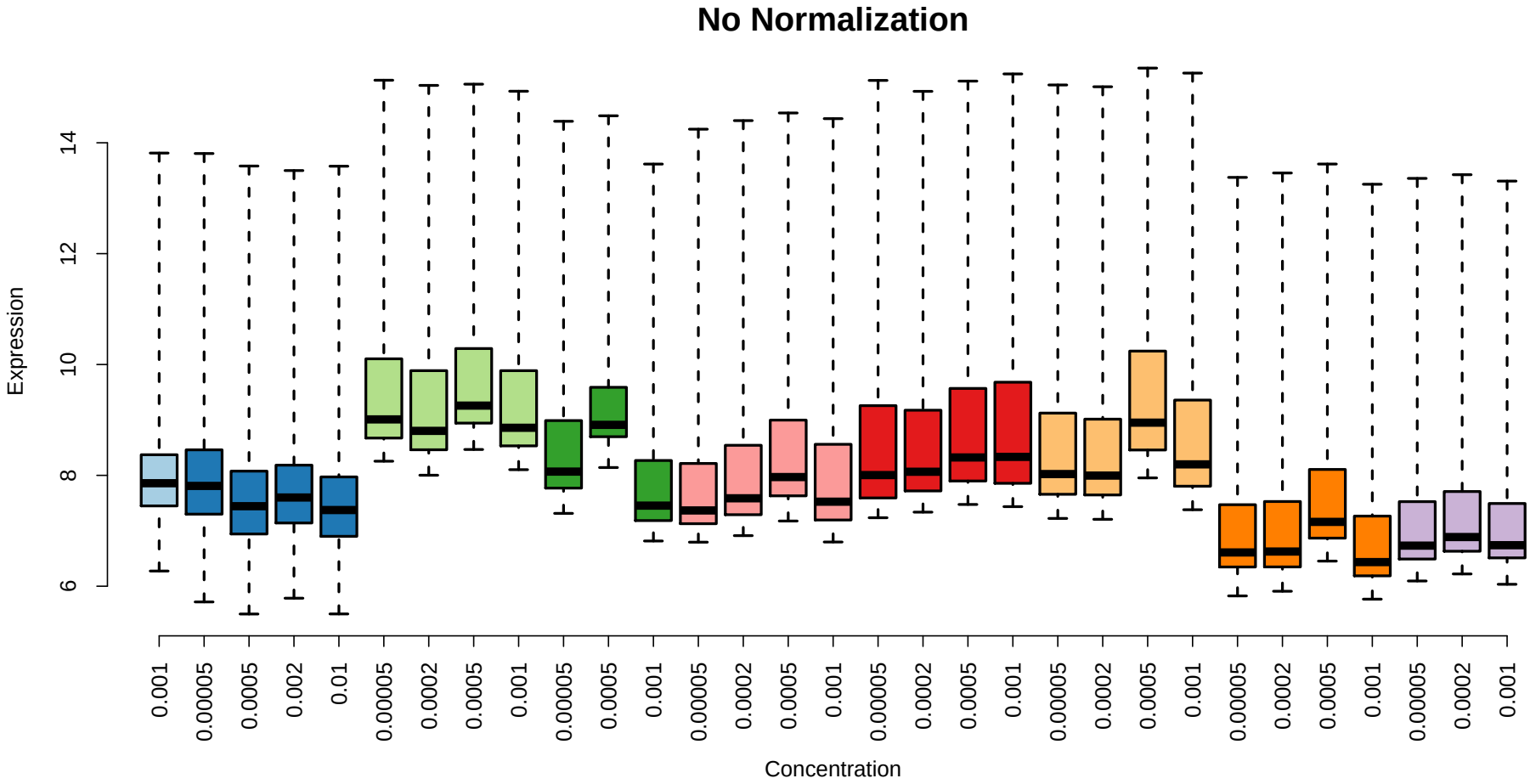
For Affymetrix arrays:

- Each component is $N(X\theta_m, \sigma_m^2)$ $(m = 1, 2)$, where

$$x_i\theta_m = \alpha_m n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{jkm} I_{ijk} + \sum_{l \in \{A,C,G,T\}} \gamma_{lm} n_{ik}^2,$$

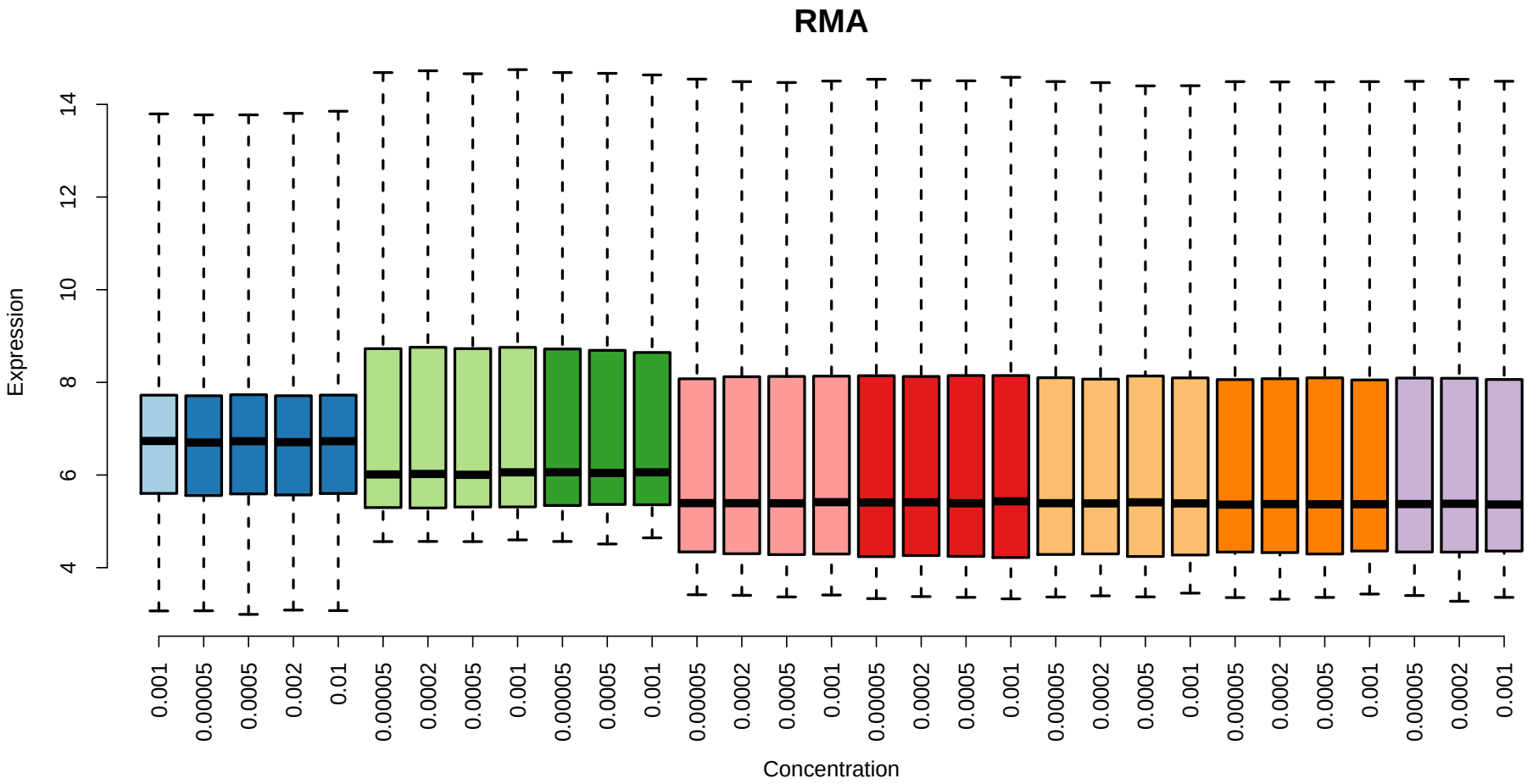
(Johnson et al., PNAS, 2006)

Batch and Design Effects



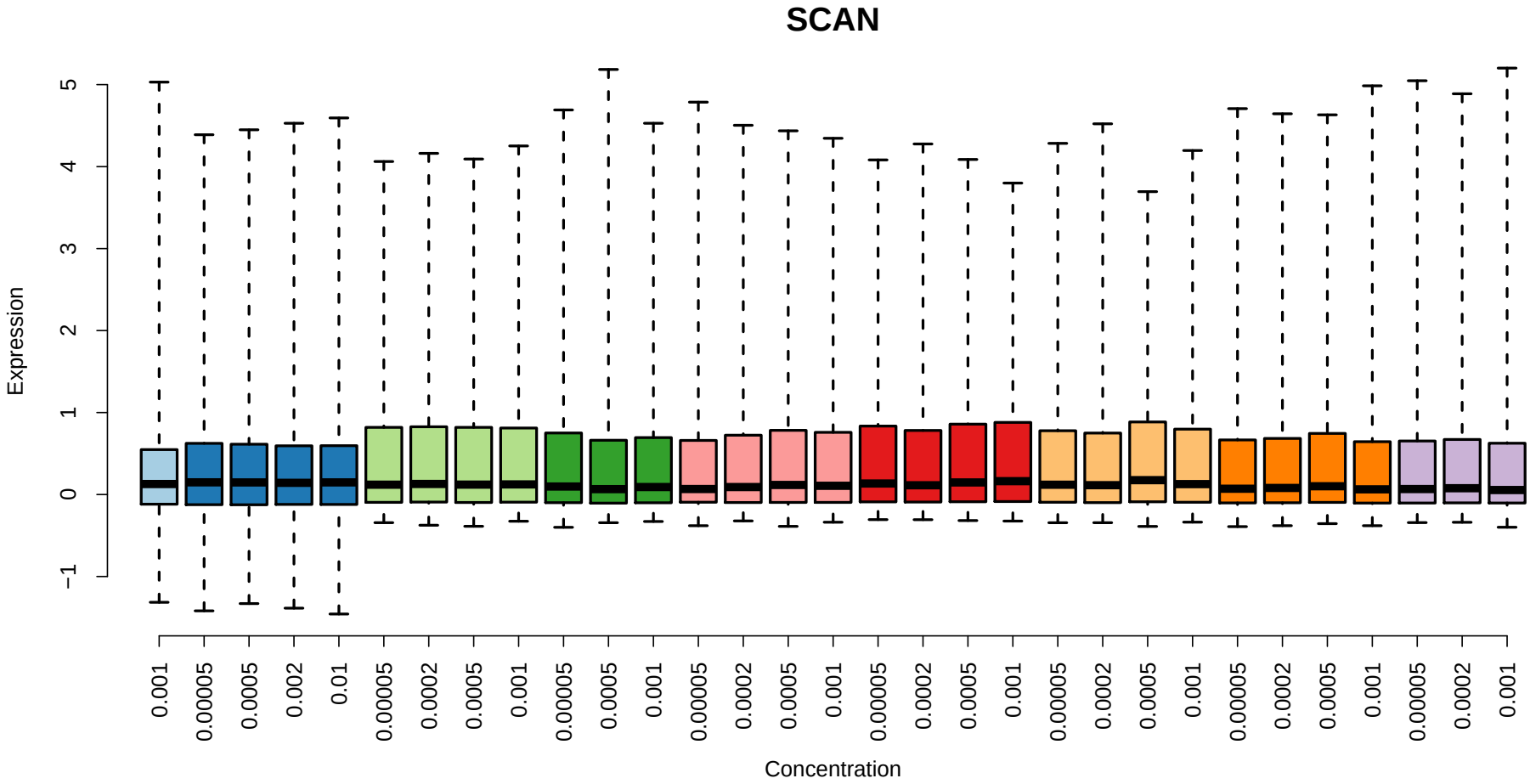
Piccolo, et al., *Genomics*, 2012

RMA



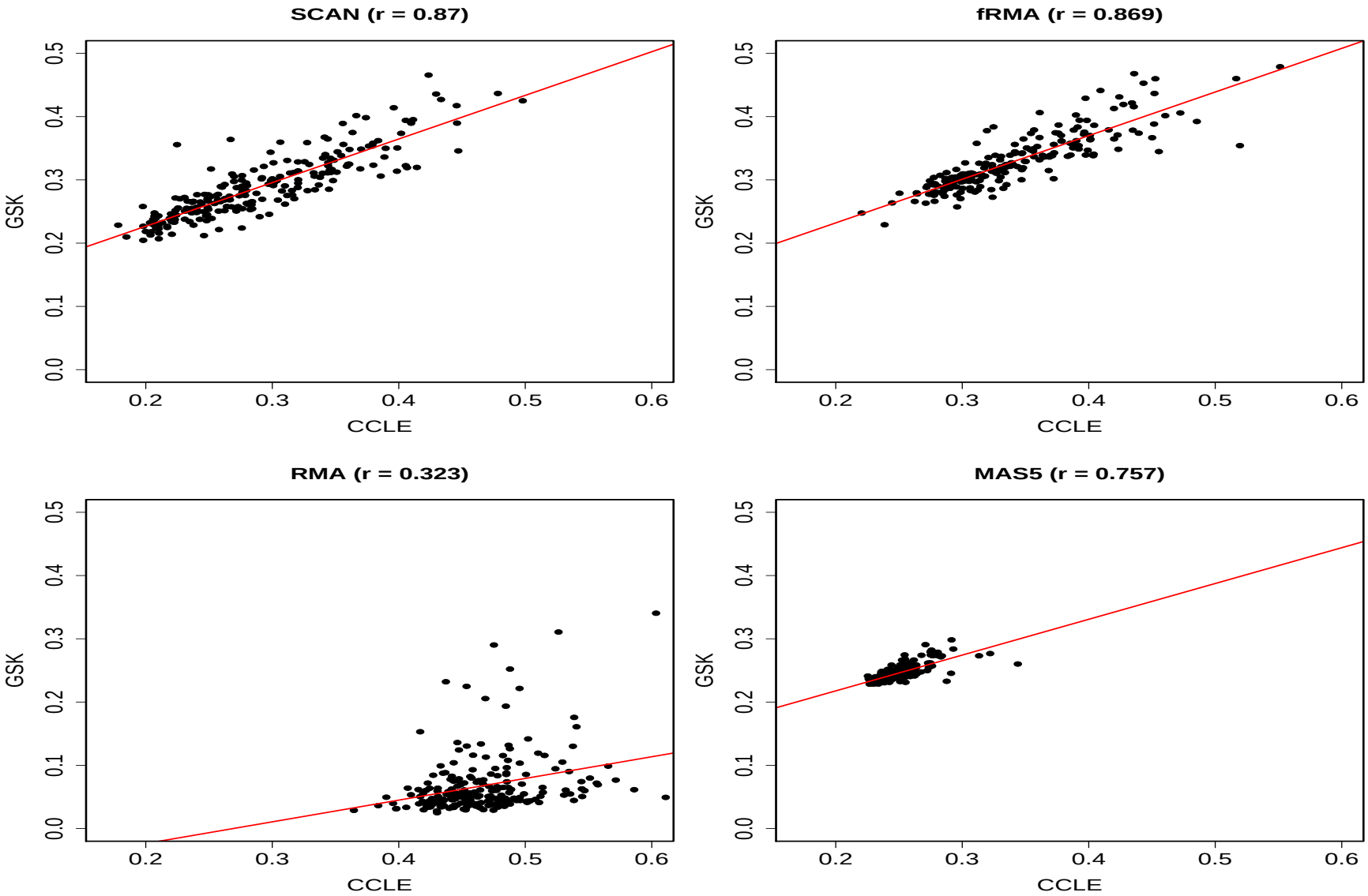
Piccolo, et al., *Genomics*, 2012

SCAN → Consistent Across Array Designs



Piccolo, et al., *Genomics*, 2012

Single Channel Array Normalization (SCAN)

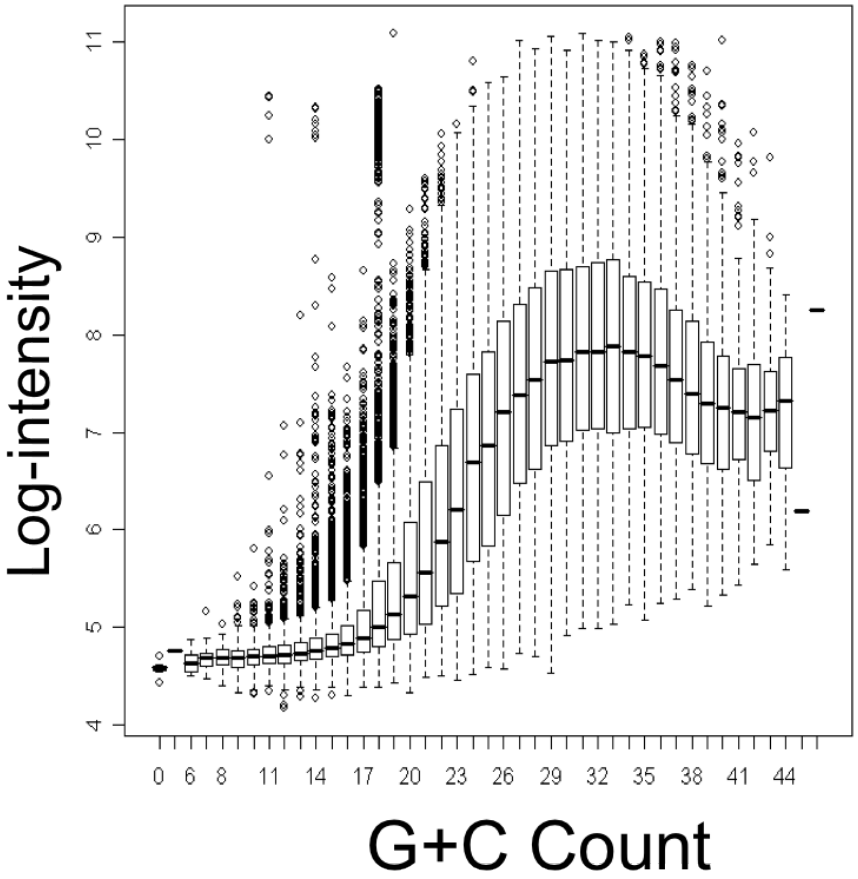


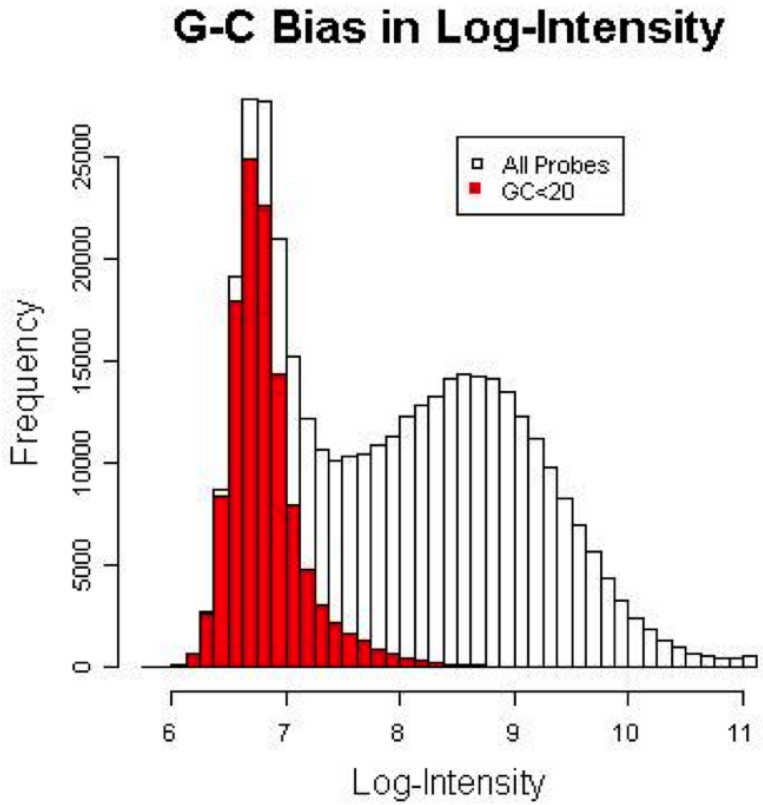
Universal Probability of Expression Code (UPC)

For two-color arrays:

1. Suppose
$$\log(\mathbf{Y_i}) = (\log(Y_{i1}), log(Y_{i2})) \sim N(\mathbf{m_k}, \Sigma_k)$$
where k is the G+C of the probe
2. Transform to mean center and remove chip and dye effects:
$$\mathbf{Z_i} = \hat{\Sigma}_k^{-1/2}(\log(\mathbf{Y_i}) - \hat{\mathbf{m}}_k)$$
(Song et al., Genome Biology, 2007)
3. Apply a simple two-component mixture model

Universal Probability of Expression Code (UPC)





Universal Probability of Expression Code (UPC)

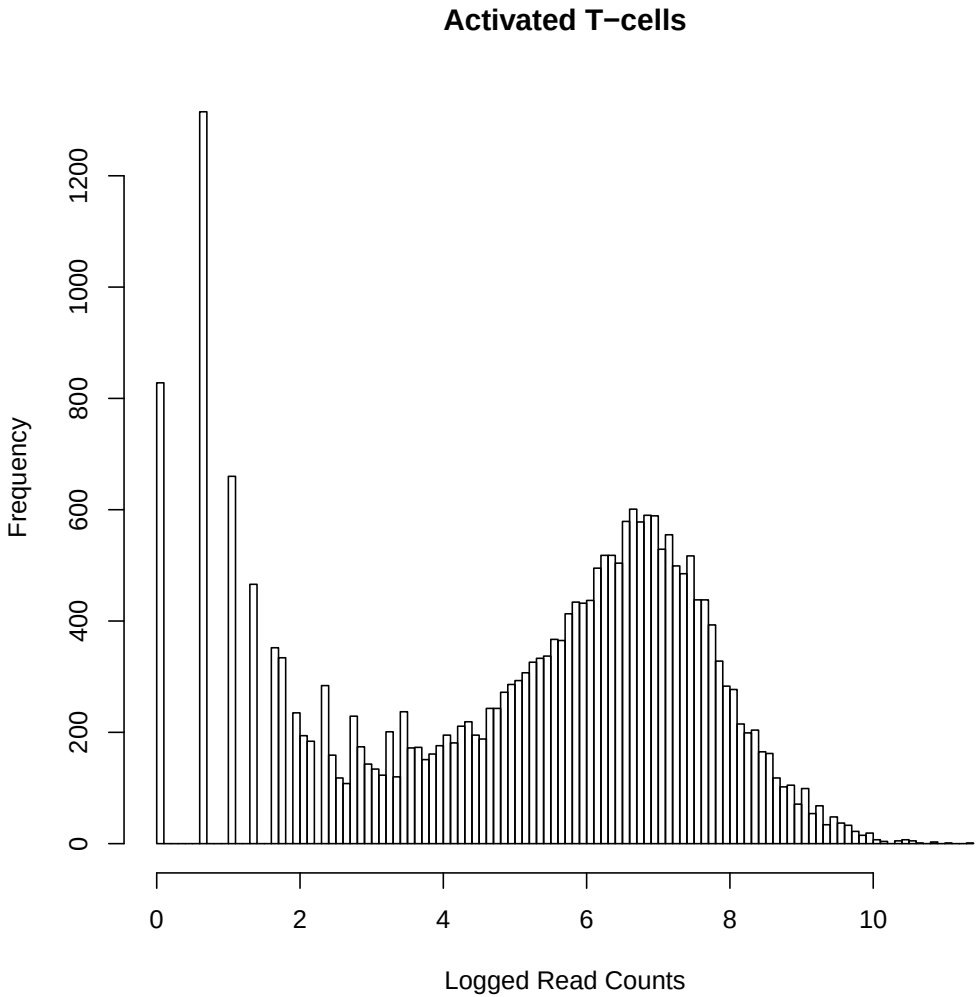
RNA-seq data:

- Mapping errors, repetitive regions
- 'Leaky' transcription
- Each component is $N(X\theta_m, \sigma_m^2)$, where

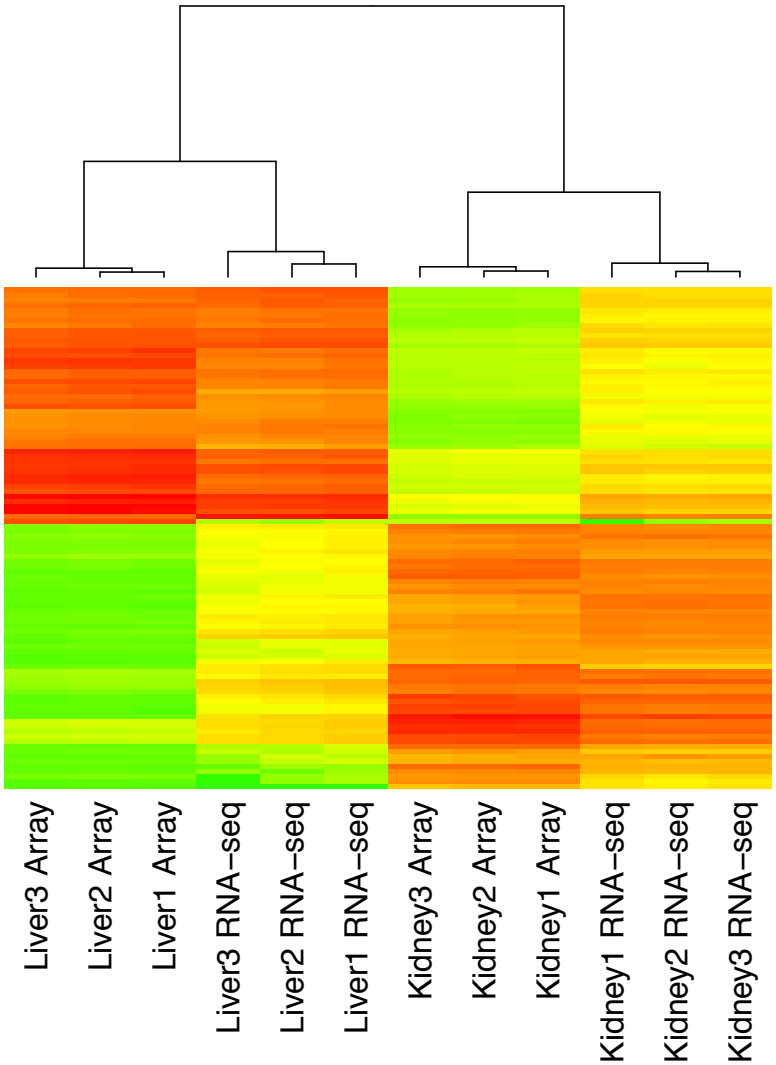
$$x_i\theta_m = \alpha_m + GC_i\beta + Len_i\gamma$$

Universal Probability of Expression Code (UPC)

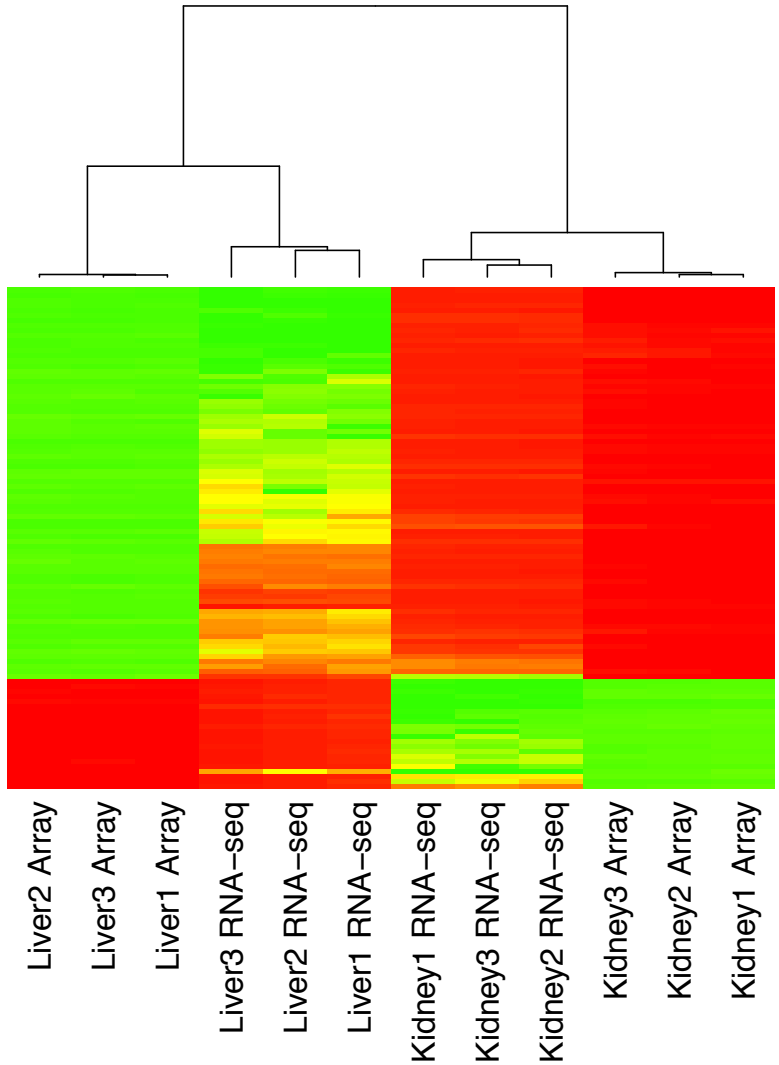
RNA-Seq Data:



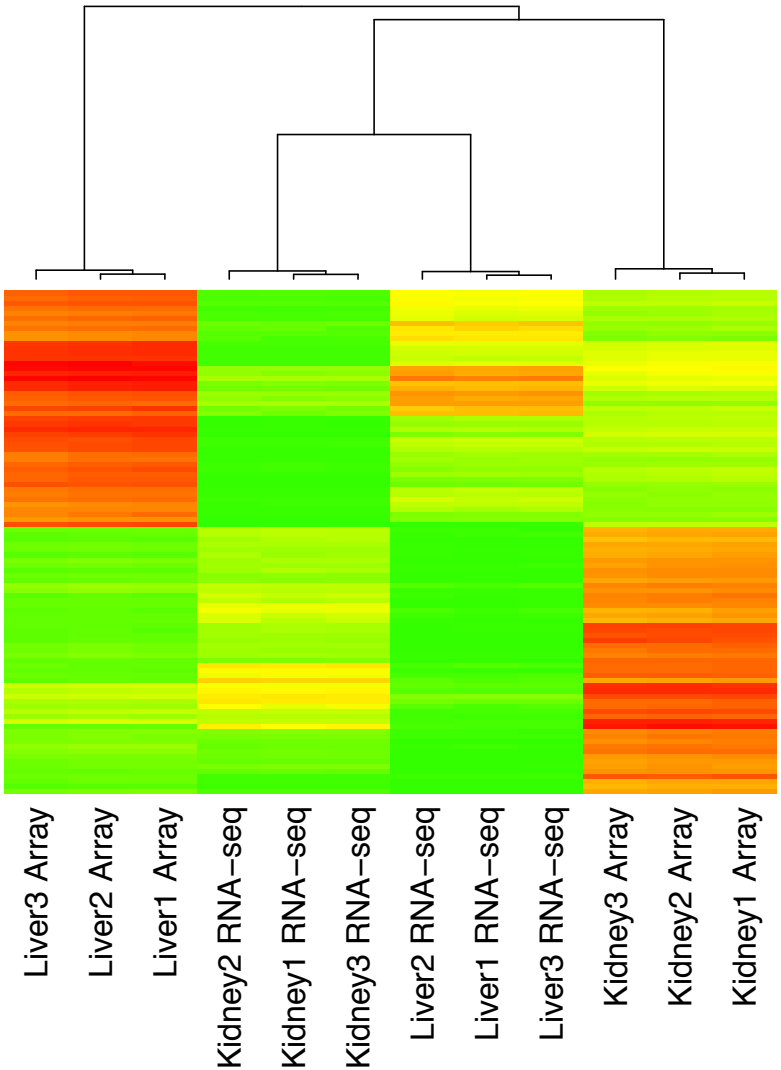
(a) Normalized Array, Read Count



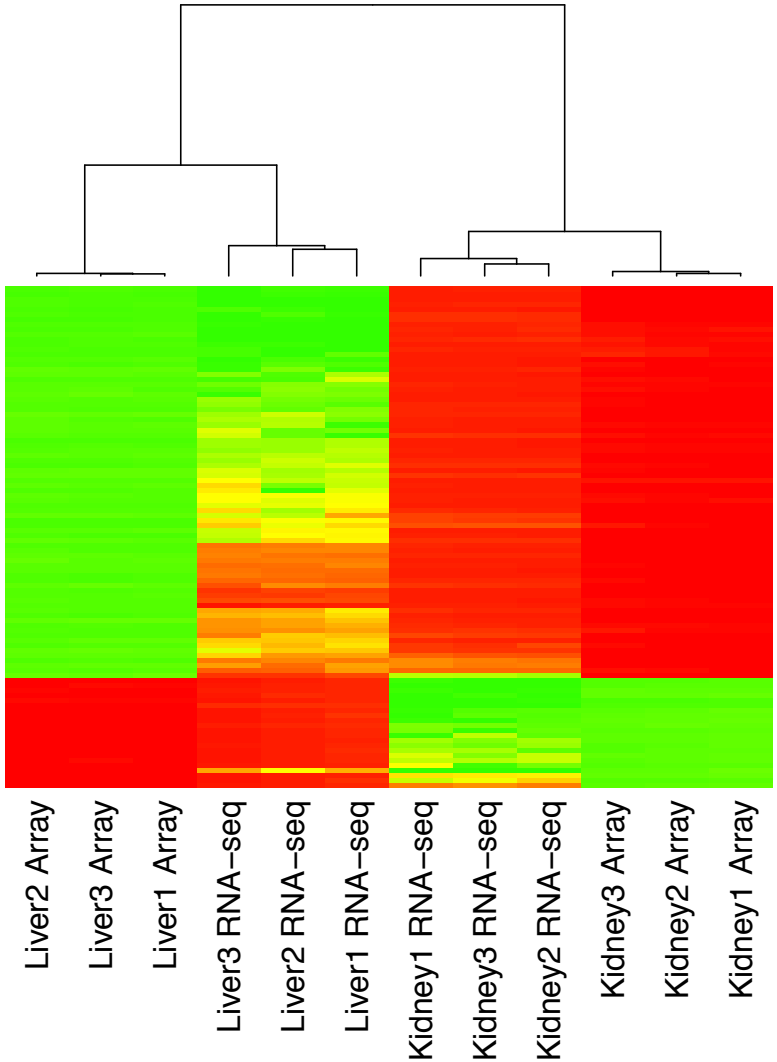
(c) UPC Array and Seq

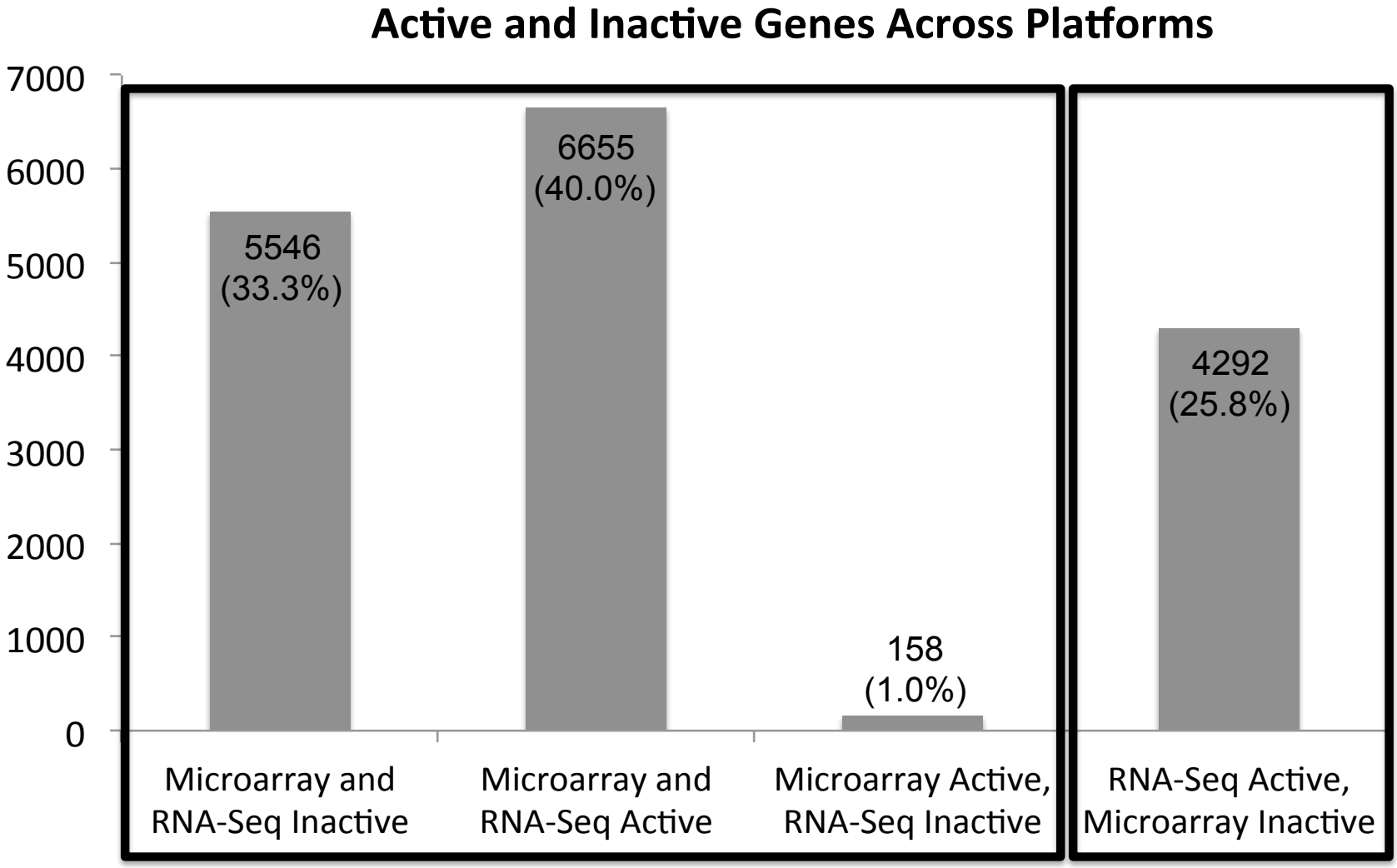


(b) Normalized Array, RPKM



(c) UPC Array and Seq

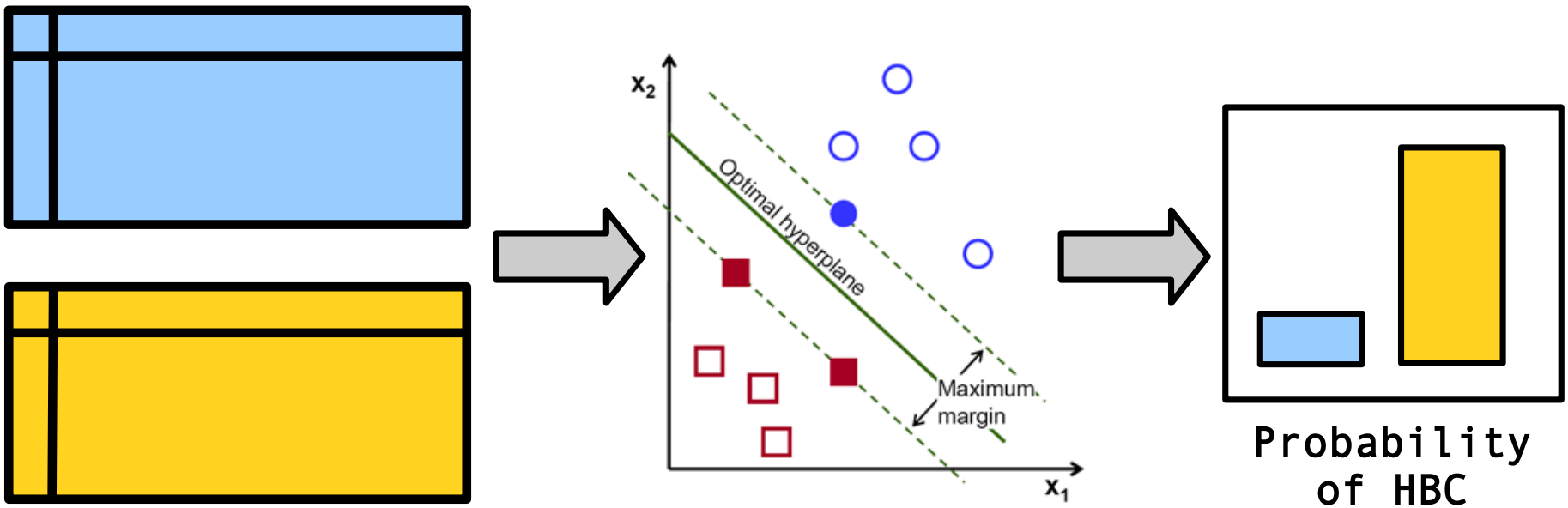




Microarray active or RNA-Seq inactive subset: 98.7% correspondence across platforms

Discordant due to increased sensitivity of RNA-Seq

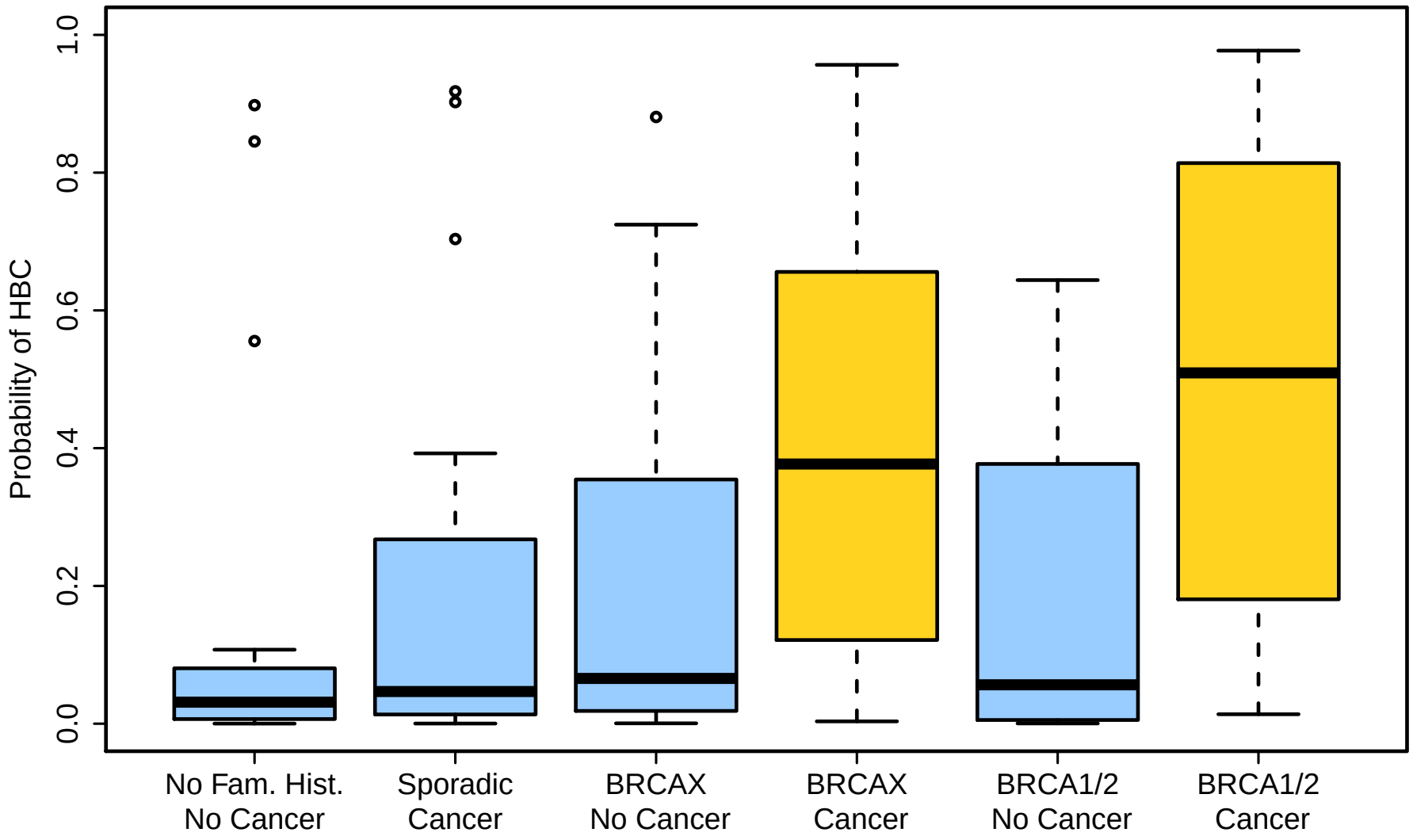
Risk Prediction



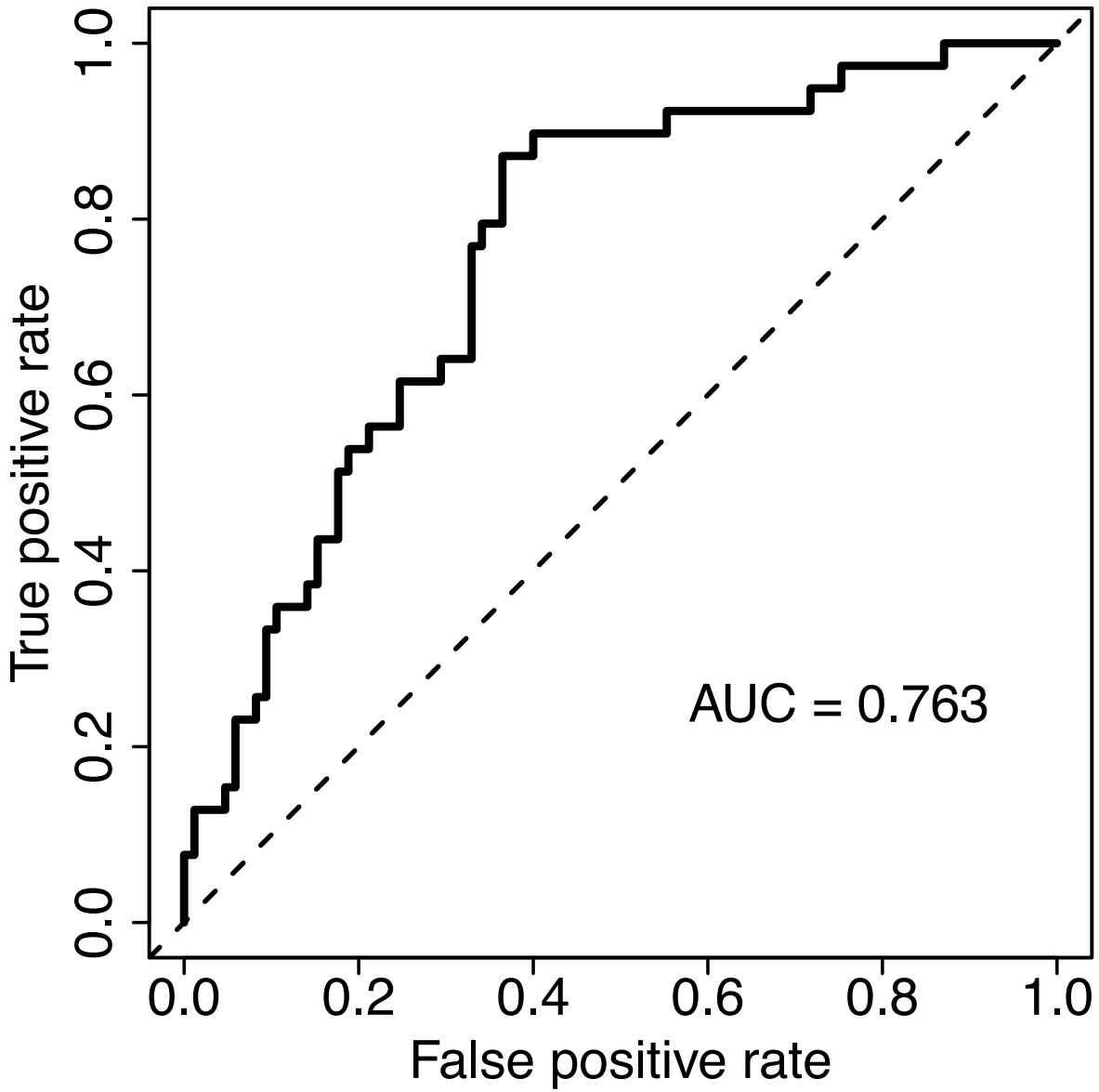
ML-Flex

Piccolo and Frey, *JMLR*, 2012

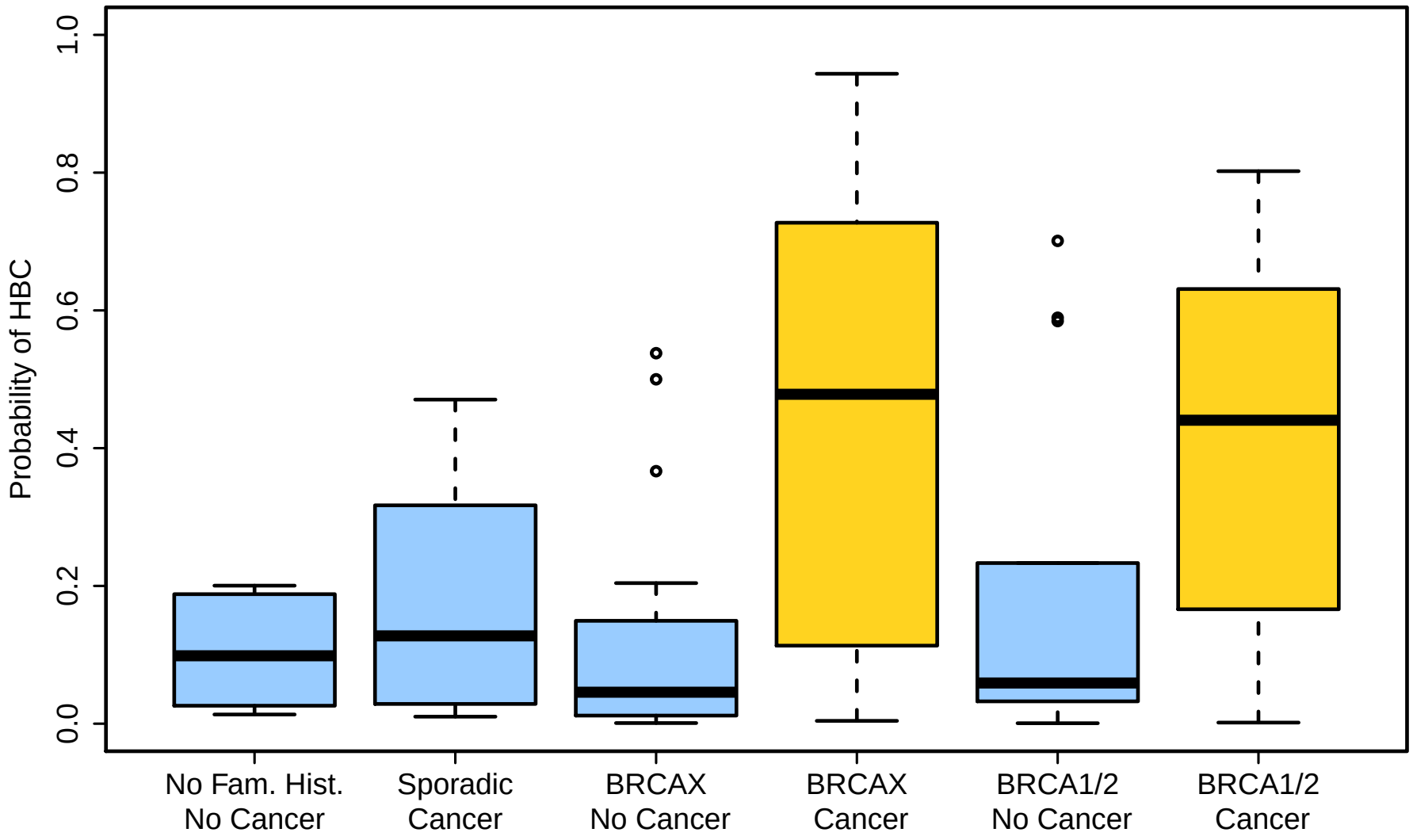
Utah – Predicted Risk



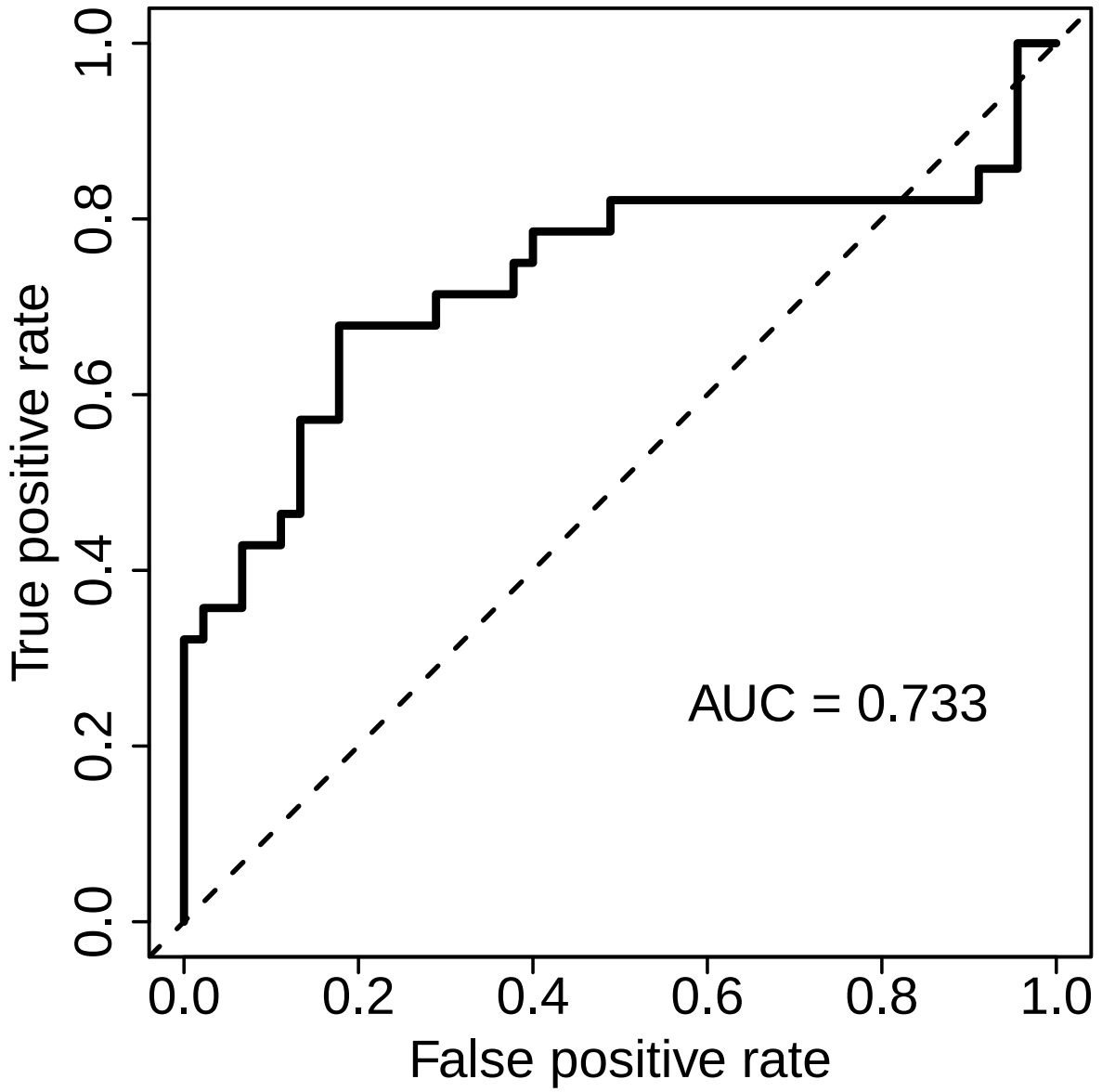
Utah — ROC Curve



Ontario – Predicted Risk



Ontario — ROC Curve

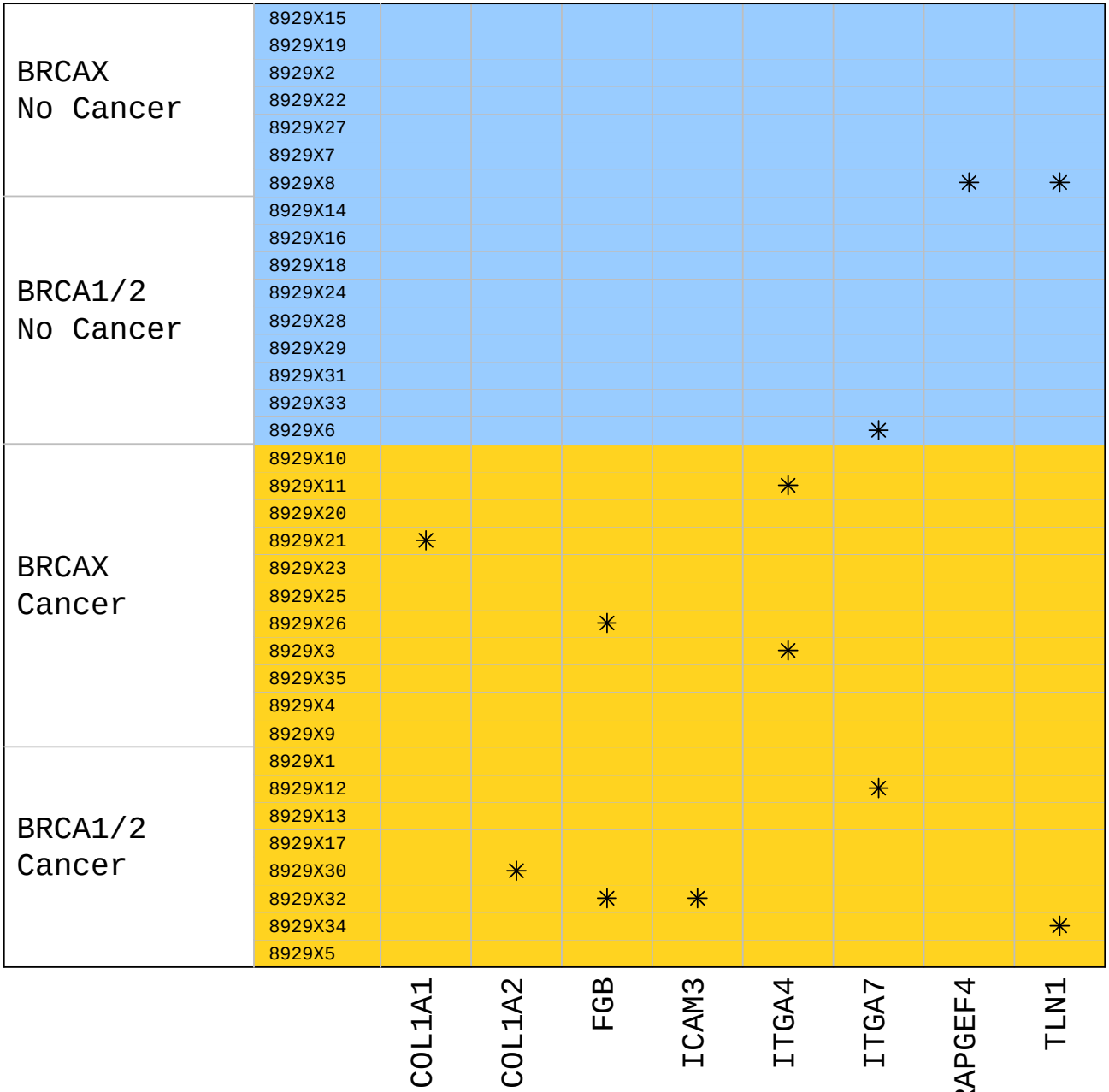


Top Pathway Results

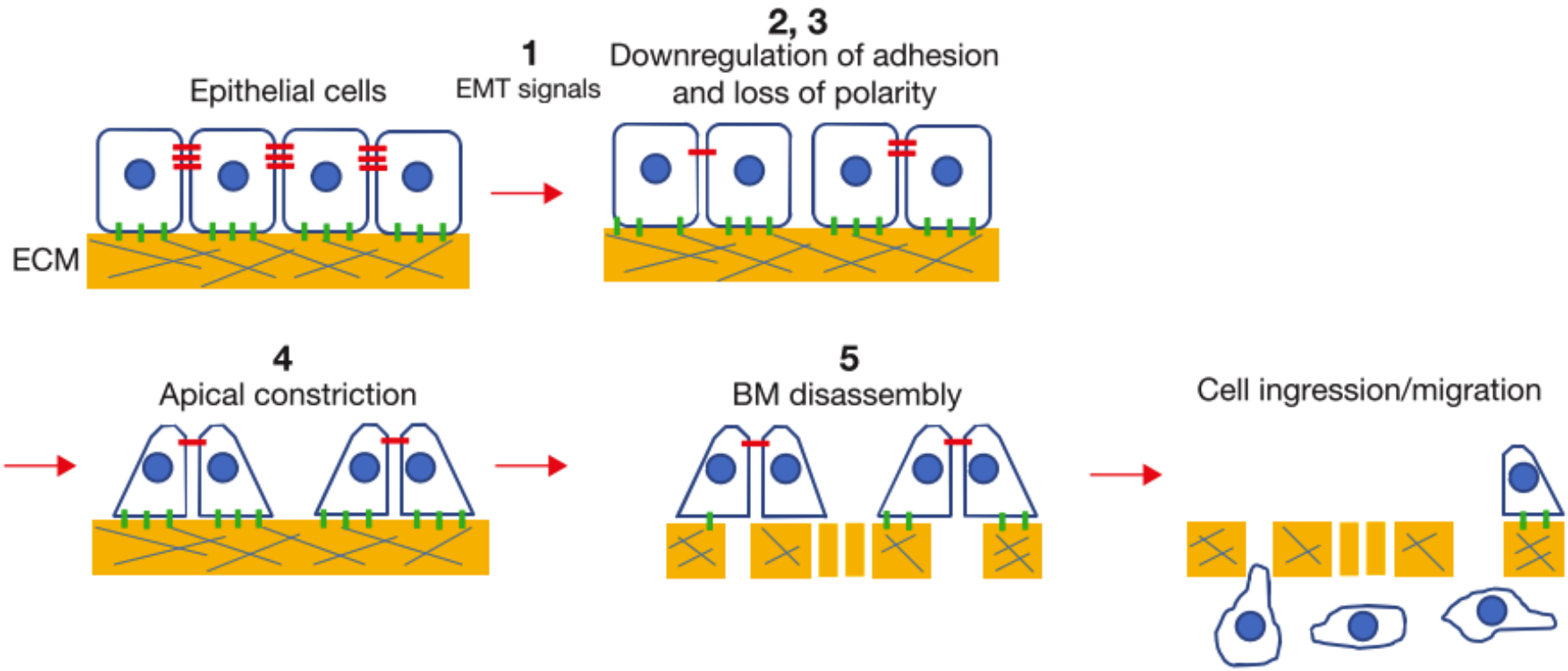
Pathway	AUC	Controls Mutated	Cancer Mutated
Integrin cell surface interactions	0.687	3/16	9/19
Cell adhesion molecules	0.682	2/16	8/19
PI3K Signaling System	0.676	4/16	10/19
Citrate/Krebs cycle	0.678	0/16	7/19
Fructose and mannose metabolism	0.668	1/16	7/19
ERBB signaling pathway	0.658	3/16	7/19

Pathways that performed well in both analyses are known to play a role in tumor development!

Integrin cell−surface interactions

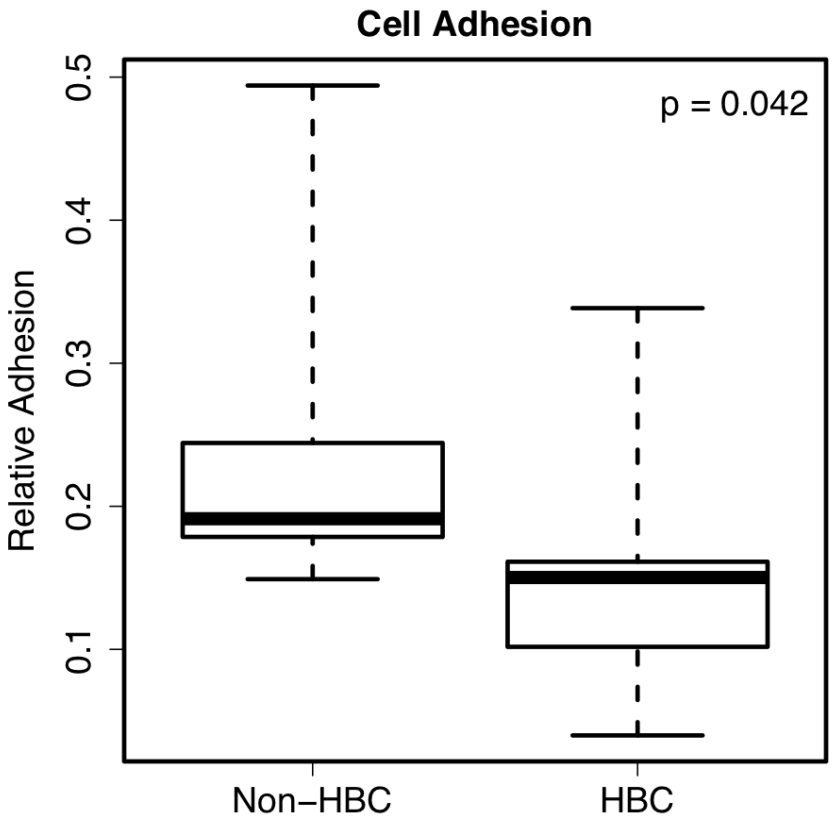


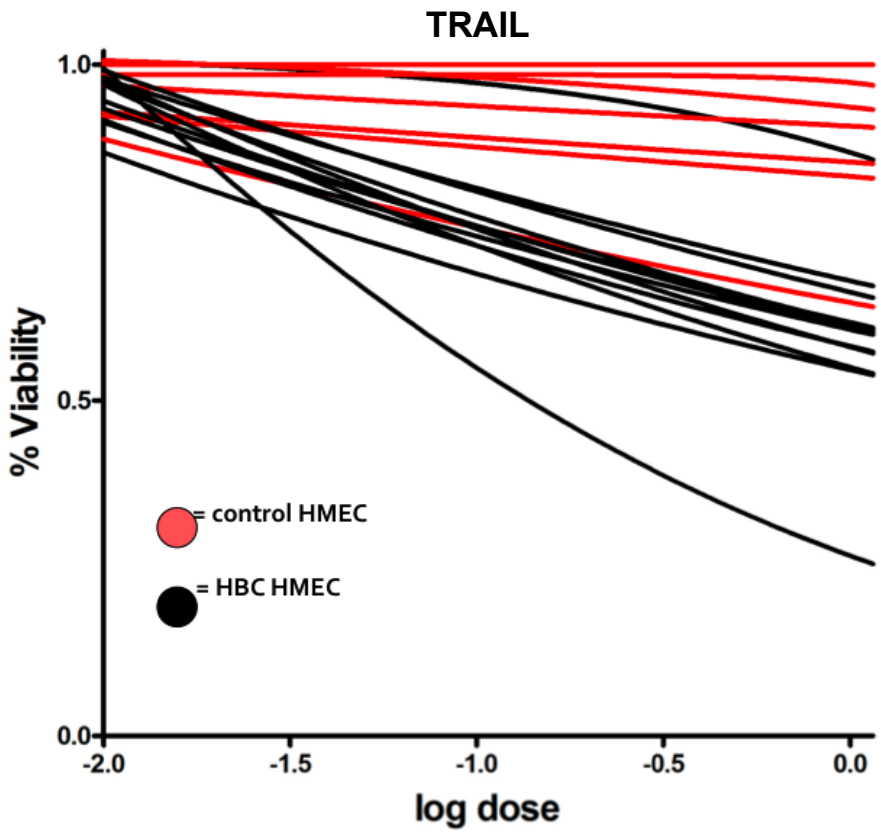
Epithelial Tissue Adhesion



Layayer and Lecuit, Nat Cel Bio, 2008

Functional Results





Acknowledgements

SCAN-UPC
Stephen Piccolo, PhD
Owen Francis
Michelle Withers

Andrea Bild, PhD
Ying Sun

Marc Lenburg
Josh Campbell

Breast Cancer Research Team:
Andrea Bild, PhD
Stephen Piccolo, PhD
Saundra Buys, MD
Theresa Werner, MD
Tom Conner
David Goldgar, PhD

Avi Spira, MD

Irene Andrulis, PhD

Funding:
NIH U01 CA164720
NIH R01 HG005692

Thank-you!